Урок №22. Системы оптического распознавания документов.

Цели: научить сканировать «бумажные» тексты и преобразовывать их в компьютерные текстовые документы с помощью систем оптического распознавания.

Требования к подготовке учащихся:

Знать/понимать: - различия в технологии распознавания текста при использовании растрового и векторного методов.

Уметь: - сканировать «бумажные» тексты и преобразовывать их в компьютерные текстовые документы с помощью систем оптического распознавания.

Использовать: - полученные знания и умения в дальнейшем.

Тип урока: практическая работа №15
Формы работы: фронтальная, индивидуальная

Ход урока:

1. Организационный момент

2. Изучение нового материала

Системы оптического распознавания символов. Системы оптического распознавания символов используются при создании электронных библиотек и архивов путем перевода книг и документов в цифровой компьютерный формат.

Сначала с помощью сканера необходимо получить изображение страницы текста в графическом формате. Далее для получения документа в текстовом формате необходимо провести распознавание текста, т. е. преобразовать элементы графического изображения в последовательность текстовых символов.

Системы оптического распознавания символов сначала определяют структуру размещения текста на странице и разбивают его на отдельные области: колонки, таблицы, изображения и т. д. Далее выделенные текстовые фрагменты графического изображения страницы разделяются на изображения отдельных символов.

Для отсканированных документов типографского качества (достаточно крупный шрифт, отсутствие плохо напечатанных символов или исправлений) распознавание символов проводится путем их сравнения с растровыми шаблонами.

Растровое изображение каждого символа последовательно накладывается на растровые шаблоны символов, хранящиеся в памяти системы оптического распознавания. Результатом распознавания является символ, шаблон которого it наибольшей степени совпадает с изображением (рис. 3.16).

Рис. 3.16. Распознаваемый символ "Б" накладывается на растровые шаблоны символов (А, Б, В и т. д.)

При распознавании документов с низким качеством печати (машинописный текст, факс и т. д.) используется векторный метод распознавания символов. В распознаваемом изображении символа выделяются геометрические примитивы (отрезки, окружности и др.) и сравниваются с векторными шаблонами символов. В результате выбирается тот символ, для которого совокупность всех геометрических примитивов и их расположение больше всего соответствует распознаваемому символу (рис. 3.17).

Рис. 3.17. Распознаваемый символ "Б" накладывается на векторные шаблоны символов (А, Б, В и т. д.)

Системы оптического распознавания символов являются "самообучающимися" (для каждого конкретного документа они создают соответствующий набор шаблонов символов), и поэтому скорость и качество распознавания многостраничного документа постепенно возрастают.

С появлением первого карманного компьютера Newton фирмы Apple в 1990 году начали создаваться системы распознавания рукописного текста. Такие системы преобразуют текст, написанный на экране карманного компьютера специальной ручкой, в текстовый компьютерный документ.

Системы оптического распознавания форм. При заполнении документов большим количеством людей (например, при сдаче выпускником школы единого государственного экзамена (ЕГЭ)) используются бланки с пустыми полями. Данные вводятся в поля печатными буквами от руки. Затем эти данные распознаются с помощью систем оптического распознавания форм и вносятся в компьютерные базы данных.

Сложность состоит в том, что необходимо распознавать символы, написанные от руки, которые довольно сильно различаются у разных людей. Кроме того, такие системы должны уметь определять, к какому полю относится распознаваемый текст.

3. Практическая работа

Задание 1. Отсканировать и преобразовать в компьютерный текстовый документ страницу учебника.

1. В операционной системе Windows запустить систему оптического распознавания документов Fine Reader.

2. В окне системы оптического распознавания щелкнуть по кнопке *Сканировать*. В появившемся окне сканера щелкнуть по кнопке *Сканировать*.

3. В окне *Проектор программы* появится отсканированное изображение текстовой страницы. Для передачи изображения систему оптического распознавания щелкнуть по кнопке *Принять*.

4. В окне системы оптического распознавания появится отсканированное изображение текстовой страницы. Для преобразования графического изображения страницы в текстовый файл щелкнуть по кнопке *Распознать*. После окончания процесса распознавания ввести команду *Файл-Сохранить как*, выбрать место сохранения, имя и тип полученного текстового файла.

5. Открыть полученный документ в текстовом редакторе и исправить возможные ошибки, допущенные в процессе распознавания.

4 Подведение итогов.

1. В чем состоят различия в технологии распознавания текста при использовании растрового и векторного методов?